
Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

1	Introduction	7
1.1	Philosophy	7
2	Background	7
2.1	Kronecker Product	7
2.2	Gradients	7
2.3	Jacobian	7
2.4	Hessian	7
2.5	Newton Optimization	7
3	Distributions	7
3.1	Distribution (Base Class)	7
3.2	Arcsine	7
3.3	Asymmetric Laplace	7
3.4	Bernoulli	7
3.5	Beta	7
3.6	Categorical	7
3.7	Cauchy	7
3.8	Chi Square	7
3.9	Conditional Model	7
3.10	Convolution	7
3.11	Data	7
3.12	Dirac Delta	7
3.13	Dirichlet	7
3.14	Exponential	7
3.15	Fisher-Snedcor (F-Distribution)	7
3.16	Gamma	7
3.17	Generator	7
3.18	Gumbel	7
3.19	Gumbel Softmax	7
3.20	Gumbel Mixture Model	7
3.21	Half Cauchy	7
3.22	Half Normal	7
3.23	Hyperbolic Secant	7
3.24	Infinite Mixture Model	7
3.25	Langevin	7
3.26	Laplace	7
3.27	Log Cauchy	7
3.28	Log Laplace	7

3.29	Log Normal	7
3.30	Logistic	7
3.31	Logit Normal	7
3.32	Mixture Model	7
3.33	Normal (Independent)	7
3.34	Normal (Multivariate)	7
3.35	Rayleigh	7
3.36	Relaxed Bernoulli	7
3.37	Student T	7
3.38	Transformed Distribution	7
3.39	Uniform	7
4	Transforms	7
4.1	Transform (Base Class)	8
4.2	Inverse Transform	8
4.3	Affine	8
4.4	Exp	8
4.5	Expm1	8
4.6	Gumbel	8
4.7	Identity	9
4.8	Log	9
4.9	Logit	9
4.10	Planar	9
4.11	Power	9
4.12	Radial	9
4.13	Reciprocal	9
4.14	Sigmoid	10
4.15	SinhArcsinh	10
4.16	Softplus	10
4.17	Softsign	10
4.18	Square	10
4.19	Tanh	10
5	Criterion	10
5.1	Divergences	10
5.1.1	Cross-Entropy	10
5.1.2	Perplexity	10
5.1.3	Forward KL Divergence	10
5.1.4	Reverse KL Divergence	11
5.1.5	Jensen-Shannon Divergence	11

5.1.6	Earth Mover's Distance	11
5.2	ELBO	11
5.3	Adversarial Loss	11
5.3.1	Adversarial Loss	13
5.3.2	GAN Loss	13
5.3.3	MMGAN Loss	13
5.3.4	WGAN Loss	13
5.3.5	LSGAN Loss	13
5.3.6	Gradient Penalty	13
5.3.7	Spectral Norm	13
6	Models	13
6.1	Regression	13
6.1.1	Linear Regression (Normal)	13
6.1.2	L1 Regression (Laplace)	13
6.1.3	Ridge Regression (Normal + Normal Prior on Weights)	13
6.1.4	Lasso Regression (Normal + Laplace Prior on Weights)	13
6.2	Classification	13
6.2.1	Logistic Regression (Bernoulli)	13
6.2.2	Bayesian Logistic Regression (Bernoulli)	13
6.2.3	Softmax Regression (Categorical)	13
6.2.4	Bernoulli Naive Bayes (Bernoulli - Bernoulli)	13
6.2.5	Gaussian Naive Bayes (Multinomial - Gaussian)	13
6.2.6	Multinomial Naive Bayes (Bernoulli - Multinomial)	13
6.2.7	Linear Discriminant Analysis (Multinomial - Shared Covariance)	13
6.2.8	Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)	13
6.3	Clustering	13
6.3.1	Gaussian Mixture Model	13
6.4	Unconstrained Matrix Factorization (Gaussian)	13
6.5	Principle Components Analysis	13
6.5.1	PCA	13
6.5.2	EM-PPCA	13
6.5.3	Variational PPCA	13
6.6	Generative Adversarial Networks	13
6.6.1	Generative Adversarial Networks	13
6.6.2	GAN Model	13
6.6.3	MMGAN Model	13
6.6.4	WGAN Model	13
6.6.5	LSGAN Model	13
6.7	Variational Auto-Encoders (TBD)	13

7	Monte Carlo	13
7.1	Monte Carlo Integration	13
7.2	Linear Congruential Generator	13
7.3	Inverse Transform Sampling	13
7.4	Box-Muller	13
7.5	Marsaglia-Bray	13
7.6	Rejection Sampling	13
7.7	MCMC: Metropolis	13
7.8	MCMC: Metropolis-Hastings	13
7.9	MCMC: Metropolis-Adjusted Langevin Algorithm (MALA)	13
7.10	MCMC: Hamiltonian Monte Carlo	13

1 Introduction

1.1 Philosophy

2 Background

2.1 Kronecker Product

2.2 Gradients

2.3 Jacobian

2.4 Hessian

2.5 Newton Optimization

3 Distributions

3.1 Distribution (Base Class)

3.2 Arcsine

3.3 Asymmetric Laplace

3.4 Bernoulli

3.5 Beta

3.6 Categorical

3.7 Cauchy

3.8 Chi Square

3.9 Conditional Model

3.10 Convolution

3.11 Data

3.12 Dirac Delta

3.13 Dirichlet

3.14 Exponential

3.15 Fisher-Snedcor (F-Distribution)

3.16 Gamma

3.17 Generator

3.18 Gumbel

3.19 Gumbel Softmax

3.20 Gumbel Mixture Model

3.21 Half Cauchy

3.22 Half Normal

3.23 Hyperbolic Secant

3.24 Infinite Mixture Model

3.25 Langevin

3.26 Laplace

3.27 Log Cauchy

4.1 Transform (Base Class)

4.2 Inverse Transform

4.3 Affine

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \mu + \sigma \cdot x \quad (1)$$

- **Inverse**

$$f^{-1}(y) = \frac{y - \mu}{\sigma} \quad (2)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \quad (3)$$

4.4 Exp

- **Parameters**

- None

- **Forward**

$$f(x) = e^x \quad (4)$$

- **Inverse**

$$f^{-1}(y) = \log y \quad (5)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (6)$$

4.5 Expm1

- **Parameters**

- None

- **Forward**

$$f(x) = e^x - 1 \quad (7)$$

- **Inverse**

$$f^{-1}(y) = \log(1 + y) \quad (8)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (9)$$

4.6 Gumbel

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \exp \left(-\exp \left(-\frac{x - \mu}{\sigma} \right) \right) \quad (10)$$

- **Inverse**

$$f^{-1}(y) = \mu - \sigma \cdot \log(-\log(y)) \quad (11)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log \left(\frac{\sigma}{-\log(y) \cdot y} \right) \quad (12)$$

4.7 Identity

4.8 Log

- **Parameters**

- None

- **Forward**

$$f(x) = \log x \quad (13)$$

- **Inverse**

$$f^{-1}(y) = \exp y \quad (14)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -y \quad (15)$$

4.9 Logit

- **Parameters**

- None

- **Forward**

$$f(x) = \log \left(\frac{x}{1-x} \right) \quad (16)$$

- **Inverse**

$$f^{-1}(y) = \frac{1}{1 + e^{-y}} \quad (17)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log (1 + e^{-y}) + \log (1 + e^y) \quad (18)$$

4.10 Planar

4.11 Power

- **Parameters**

- Power p

- **Forward**

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \quad (19)$$

- **Inverse**

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \quad (20)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1 \right) \cdot \log (x \cdot p + 1) & \text{otherwise} \end{cases} \quad (21)$$

4.12 Radial

4.13 Reciprocal

- **Parameters**

- None

- **Forward**

$$f(x) = 1/x \quad (22)$$

- **Inverse**

$$f^{-1}(y) = 1/y \quad (23)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -2 \cdot \log |x| \quad (24)$$

4.14 Sigmoid

- **Parameters**

- None

- **Forward**

$$f(x) = \frac{1}{1 + e^{-x}} \quad (25)$$

- **Inverse**

$$f^{-1}(y) = \log \left(\frac{y}{1 - y} \right) \quad (26)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log(1 + e^{-x}) - \log(1 + e^x) \quad (27)$$

4.15 SinhArcsinh

4.16 Softplus

4.17 Softsign

4.18 Square

4.19 Tanh

5 Criterion

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the parameters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p . In addition, we provide the Monte Carlo approximation.

5.1 Divergences

5.1.1 Cross-Entropy

$$\begin{aligned} H(p, q) &= - \int p(x) \log q(x) dx \\ &= - \frac{1}{n} \sum_{x \sim p} \log q(x) \end{aligned} \quad (28)$$

5.1.2 Perplexity

$$\begin{aligned} H(p, q) &= \exp \left(- \int p(x) \log q(x) dx \right) \\ &= \exp \left(- \frac{1}{n} \sum_{x \sim p} \log q(x) \right) \end{aligned} \quad (29)$$

5.1.3 Forward KL Divergence

$$\begin{aligned} H(p, q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \frac{1}{n} \sum_{x \sim p} \log \frac{p(x)}{q(x)} \end{aligned} \quad (30)$$

5.1.4 Reverse KL Divergence

$$\begin{aligned} H(p, q) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \frac{1}{n} \sum_{x \sim q} \log \frac{q(x)}{p(x)} \end{aligned} \tag{31}$$

5.1.5 Jensen-Shannon Divergence

5.1.6 Earth Mover's Distance

5.2 ELBO

5.3 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q . More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q .

5.3.1 Adversarial Loss

5.3.2 GAN Loss

5.3.3 MMGAN Loss

5.3.4 WGAN Loss

5.3.5 LSGAN Loss

5.3.6 Gradient Penalty

5.3.7 Spectral Norm

6 Models

6.1 Regression

6.1.1 Linear Regression (Normal)

6.1.2 L1 Regression (Laplace)

6.1.3 Ridge Regression (Normal + Normal Prior on Weights)

6.1.4 Lasso Regression (Normal + Laplace Prior on Weights)

6.2 Classification

6.2.1 Logistic Regression (Bernoulli)

6.2.2 Bayesian Logistic Regression (Bernoulli)

6.2.3 Softmax Regression (Categorical)

6.2.4 Bernoulli Naive Bayes (Bernoulli - Bernoulli)

6.2.5 Gaussian Naive Bayes (Multinomial - Gaussian)

6.2.6 Multinomial Naive Bayes (Bernoulli - Multinomial)

6.2.7 Linear Discriminant Analysis (Multinomial - Shared Covariance)

6.2.8 Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)

6.3 Clustering

6.3.1 Gaussian Mixture Model

6.4 Unconstrained Matrix Factorization (Gaussian)

6.5 Principle Components Analysis

6.5.1 PCA

6.5.2 EM-PPCA

6.5.3 Variational PPCA

6.6 Generative Adversarial Networks

6.6.1 Generative Adversarial Networks

6.6.2 GAN Model

6.6.3 MMGAN Model

6.6.4 WGAN Model

6.6.5 LSGAN Model

6.7 Variational Auto-Encoders (TBD)

13

7 Monte Carlo

7.1 Monte Carlo Integration